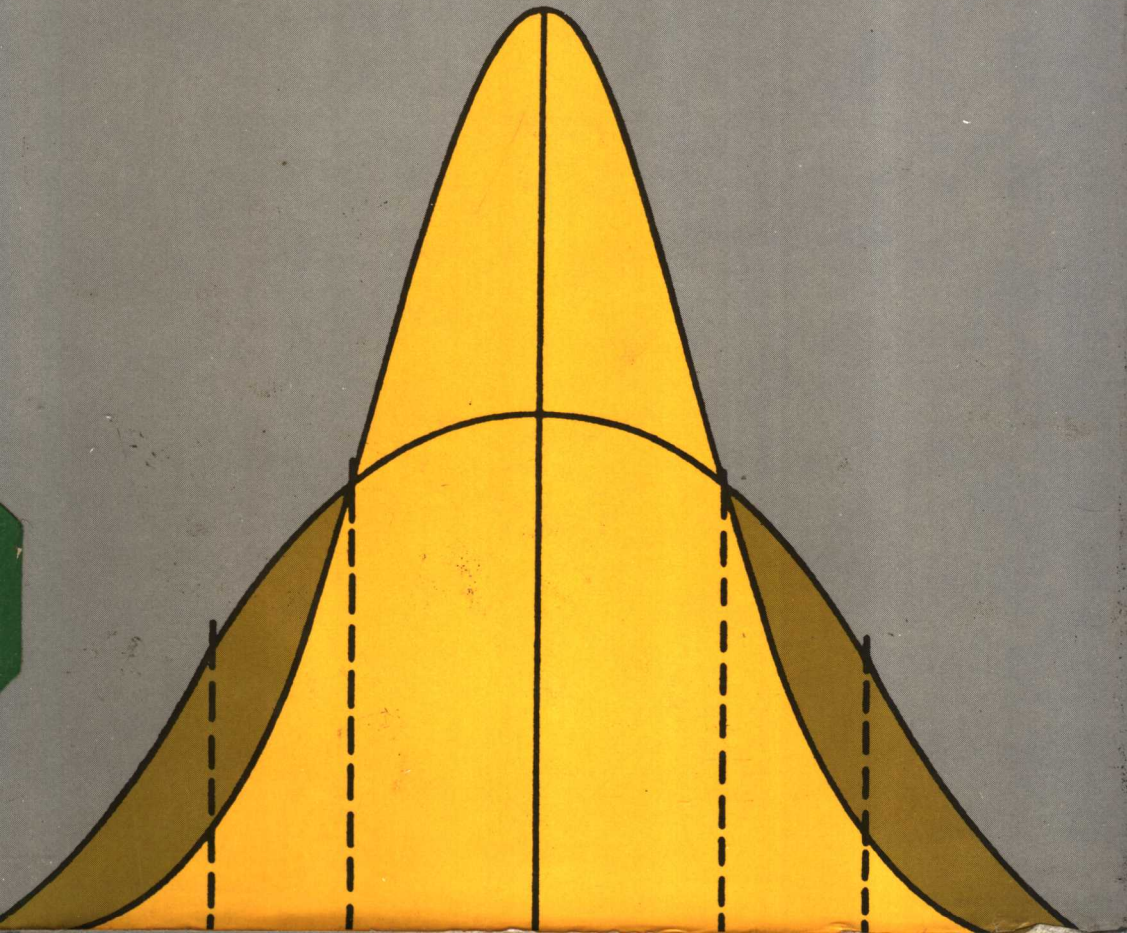
BIOSTATISTICS

SECOND EDITION

Alvin E. Lewis



BIOSTATISTICS

SECOND EDITION

Alvin E. Lewis

VNR VAN NOSTRAND REINHOLD COMPANY
NEW YORK CINCINNATI TORONTO LONDON MELBOURNE

Library of Congress Catalog Card Number: 83-16744
ISBN: 0-442-25954-9

Manufactured in the United States of America

Published by Van Nostrand Reinhold Company Inc.
135 West 50th Street
New York, New York 10020

Van Nostrand Reinhold Company Limited
Molly Millars Lane
Wokingham, Berkshire RG11 2PY, England

Van Nostrand Reinhold
480 Latrobe Street
Melbourne, Victoria 3000, Australia

Macmillan of Canada
Division of Gage Publishing Limited
164 Commander Boulevard
Agincourt, Ontario M1S 3C7, Canada

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging in Publication Data

Lewis, Alvin Edward, 1916–
Biostatistics.

Bibliography: p.
Includes index.
1. Biometry. I. Title.
QH323.5.L47 1984 574'.072 83-16744
ISBN 0-442-25954-9

CONSULTING EDITOR'S STATEMENT TO THE FIRST EDITION

One of the difficulties in teaching contemporary biology is to find a statistics text that joins biology to mathematics in a marriage of love, rather than of convenience. Dr. Lewis has been outstandingly successful in this difficult task for two reasons. First, he has based the book on sound mathematical principles and, second, he has copiously illustrated these principles with the type of biological data that belongs to today. The book gets right down to fundamentals by discussing the relation between probability and randomization, and then deals with random selection and the description of data. A satisfactorily brief chapter on the "normal" curve leads first to sampling and the universe of discourse, and then to a consideration of the null hypothesis. The student is then taken smoothly through regression, correlation, enumeration statistics, Chi square, and the analysis of variance to a thorough discussion of tolerance limits, including a chapter, rare in biology statistics books, on quality control and its application to biological problems. The book comes to an end with discussions of alternatives to the null hypothesis, sequential analysis, and a chapter on nonparametric statistics. Each of the sixteen chapters has at its end, for ready reference, a summary of the chapter. The appendix has an unusually full set of those tables necessary to the working statistician.

This book is what I have long wanted for my own course in biostatistics so that I welcome it as an admirable newcomer to the REINHOLD BOOKS IN THE BIOLOGICAL SCIENCES, not only as an editor, but also as a teacher.

PETER GRAY

PREFACE TO THE SECOND EDITION

When the first edition of this book was written over 15 years ago, several topics were included that seemed useful and important for biological investigations. As the years passed there was little need to use them either directly or in consultation, and our view of their importance has been drastically revised. At the same time topics we then regarded as less useful we have now come to consider as essential.

In spite of these misguided enthusiasms the original edition and its Spanish translation continue to be used. It is gratifying then to have an opportunity to make important improvements in this second edition. This introductory text should now be more directly helpful in the evaluation of medical and biological data and in the planning of investigations.

For all that has been done by colleagues, friends, and loved ones to help me, my gratitude is undiminished by the passage of time. My effort here to provide a lucid introduction to biological statistics is in part an expression of my appreciation.

A. E. LEWIS

PREFACE TO FIRST EDITION

If this were the first book ever written on biological statistics, this author's task would have been a simple one. There are, to be sure, a remarkable number of texts on this subject as well as on the subject of statistics in general. These vary from simple formularies with instructions covering typical applications to sophisticated treatises on statistical mathematics. To a degree these would appear to match the particular needs of specific groups of students with varying levels of mathematical skill. However, in spite of the wealth of available choices, a substantial gap remains which, hopefully, this book will fill.

This book is designed for students in biology and medicine who have reached the stage where they are ready to judge data and to begin their own investigations and experiments. It is assumed that these readers are able to follow ordinary algebraic manipulations. A background that includes calculus would guarantee an adequate mastery of the necessary algebra, but calculus is not required for this text. On the other hand, the student who has had no algebra beyond the high school level and has avoided using even this will find some portions of this book a little tedious.

Relatively little mathematical skill is demanded of the reader, but at the same time the aim here is to provide sufficient insight into the processes of statistical analysis to use them intelligently. The book is intended to take the student beyond the usual introductory expositions of the t-test, χ^2, regression, and correlation. Particular care has been given to the exposition of the analysis of variance; even if this is not included in a one semester course, the student may need it later in his career for the design of experiments. Other items presented that are not usually included in these elementary courses are quality control as applied to biological and clinical investigations, some

nonparametric methods, elements of sequential analysis, and hypotheses testing with particular emphasis on their relationship to determinations of sample size. This text should continue to serve the student long after the completion of formal course work.

Wherever possible the algebraic developments have been linked to situations. Since these expositions rely heavily on intuitive reasoning, they are often lacking in either mathematical elegance or rigor. However, this book is not designed for mathematicians, but for biologists and clinicians who otherwise would not find this material in a form easily intelligible to them.

For whatever merit this book may possess I am indebted beyond measure to those who have taught me. My interest in this subject was initiated by the late F. W. Weymouth who along with John Field, II and Victor E. Hall guided my first struggles with biological measurements in the Physiology Department at Stanford University. For the courage to look further into the theoretical basis of measurement, I must thank my former colleagues at U.C.L.A., Moses A. Greenfield and Amos Norman. Their friendly and informal discussions ranging from games theory to mutation rates made me aware of the role of probability theory in areas outside the confines of my own special interests.

Most of this book was written during the time that I was responsible for the operations of the clinical laboratories and for the training of residents in Clinical Pathology at the Mount Zion Hospital and Medical Center in San Francisco. I shall always gratefully remember the challenges and kindly encouragement of my colleagues and students at this institution.

I am indebted to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., Cambridge, to Dr. Frank Yates, F.R.S., Rothamsted, and to Messrs. Oliver & Boyd Ltd., for permission to reprint Tables Nos. A2, A3, A4, A7, and A8 from their book *Statistical Tables for Biological, Agricultural and Medical Research.* I am indebted to the RAND Corporation of Santa Monica, California for permission to reprint the material in Table No. A1 from their publication, *A Million Random Digits.* I am also indebted to D. L. Burkholder, Editor of *The Annals of Mathematical Statistics* for permission to reprint Table A5 from "Tabulated Values for Rank Correlation," by E. G. Olds, appearing in Volume IX of the Annals for 1938.

Just as actors need a producer and a stage to reach their audience, an author needs a publisher to provide the management of the

complex operations of the publishing industry. The staff of the Reinhold Publishing Corporation has been most helpful with this venture from its inception. I am particularly grateful to Mr. Leonard Roberts, Editor, for his stimulating encouragement and to Mrs. Cynthia Harris, Copy Editorial Supervisor, for her patience and for her meticulous review of the manuscript.

Finally, I must pay tribute to my wife and to my daughters. Without their patience, understanding, and encouragement this project would have ended with Chapter I. This book is dedicated, then, to Doris, Joan, and Elizabeth from an affectionate husband and father as a small measure of compensation of the time together that we have lost forever.

April, 1966 ALVIN E. LEWIS

CONTENTS

1
INTRODUCTION

While the mathematical theory of statistical methods is not usually part of the standard equipment of students in biology, statistical reasoning and techniques can be mastered and applied by anyone with a fair grasp of algebra. Nothing would be gained by suggesting that these methods are easy. They do require thought and application, but their inherent interest and obvious utility can make them reasonably pleasant and satisfying.

The subleties of statistical reasoning often escape the student in his first encounter with statistical methods. Obtaining the answer to textbook problems is only the first step in mastering statistics. Mystique has no place in science, but in an elementary exposition of this kind intuitive understanding will have to be substituted for several years of mathematical preparation. The purpose of this text is to aid the student in acquiring a useful mastery of the subject by drawing wherever possible on the commonsense experience of daily judgements.

Consider first the well-known average and some of its ordinary uses. For example, suppose we need to order a ten day supply of food for 100 experimental animals. If the average requirement per animal for ten days is known, all we have to do is multiply this amount by 100. Even though some animals eat much more than the average, we do not need much statistical knowledge to feel reasonably sure that the large eaters will be balanced by an almost equal number of small eaters. On the other hand, if instead of 100 animals, we have to feed only one or two, we would not be surprised if the average amount of food turned out to be either excessive or inadequate. This obvious example gives an intuitive basis for making a few useful general statements.

First, there is the average itself. If the average food supply should be given in a handbook or manual, we would assume that the value

reported is a summary of some comparable experience with a large group of animals, although this assumption might not be made consciously. This common average is more properly called an *arithmetic mean* in the language of statistics. As shall be pointed out later, there is also a *geometric* and a *harmonic mean*. Each of these is quite different from the arithmetic mean. Generally, when the term *mean* is used without a modifier, the arithmetic mean or average is implied.

The mean, then, regardless of type, summarizes in a single number many individual values. A summarizing value of this kind is necessary, because the human mind is unable to grasp multiple impressions simultaneously. Again, we intuitively or unconsciously assume that a mean or average is significant and useful only when it summarizes a large number of values. We assume that the average food requirement in the handbook was obtained on the basis of a large number of animals and do not expect to apply this value precisely unless we too are dealing with a large number of comparable animals.

So far, most of the discussion has been an elaboration of the obvious; this is necessary perhaps, but fairly obvious all the same. One assumption was noted, however, which may not be obvious and in spite of our intentions may not always be true. This assumption is that each animal eating a certain amount more than the average would be matched by another in the group eating almost the same amount less than the average. In other words, the animals could probably be paired off so that the average for each pair would equal the average for the group. This implies that the frequencies of each value are symmetrically distributed on either side of the mean.

Another notion, which is readily accepted, concerns the range of food requirements. The range of requirements is the difference between the largest and the smallest values in the group. Without any intensive thought, or recall of past experience, most of us would readily agree that extremes in food requirement are unusual. That is, most of the animals would have requirements not too different from the mean. The extremes at either end of the scale represent a small minority.

So far, the discussion has been an intuitive, qualitative description of a common or *normal frequency distribution*. Symmetrical frequency distributions with the majority clustered about the mean occur in numerous instances. In the study of statistical analysis these concepts will be used quantitatively.

Up to this point, animal dietary requirements have been used as an example of the average and the distribution of values of a *continuous variable.* There are many kinds of measurements that fall into this category. These are measurements that increase by vanishingly small amounts, the smallness being limited by one's ability to discriminate correspondingly fine differences. For example, if we were to take 1000 people, all of them weighing between 150 and 151 pounds, we could, if the scales were sufficiently sensitive, arrange them in order of increasing weight. Obviously, in order to do this, we would need a scale that could discriminate differences smaller than 1/1000 of a pound. Theoretically, with an unlimited population to draw from, we could take 1000 people weighing between 150.002 and 150.003 pounds and also arrange them in order of weight if the scales were sensitive to less than 1/1,000,000 of a pound. Needless to say, in real life there would never be any occasion to carry these measurements to such hair splitting accuracy. Nevertheless, a mean value of, say, 150.01 pounds has conceptual reality.

On the other hand, if a statistician states that the average family has 2.3 children, we balk at the image of three tenths of a child. We do not for a moment deny the utility of this mean for certain economic purposes, but we can immediately perceive that another class of values is involved. These are called *discontinuous variables.* They are obtained by counting or simple enumeration rather than by measuring against a scale of some kind. In genetics we *count* progeny with distinctive characteristics; in studying epidemics, we *count* cases; in bacteriology we *count* organisms. In all of these examples the units are indivisible. The count moves up discontinuous steps instead of rolling up a continuous slope.

Our commonsense, intuitive grasp of chance and counting again comes to our aid in understanding the statistical behavior of discontinuous variates. Indeed, our grasp of probability theory, which is needed in order to make statistical analysis useful, has its roots in our intuitive notions about such variates. A simple example will suffice for now, and although superficially it may appear trivial, the same example will be explored in considerable depth in a later chapter.

Suppose we take 100 coins, shake them in a box, and see how many turn up heads and how many tails. Without actually doing this experiment, we expect close to 50 heads and 50 tails. Remember that in performing this task, we would find the number of heads by

counting, not by measuring. There might be 47 heads instead of the expected 50 but obviously there will not be 47.4 heads. The important intuitive point here is that we *expect* close to 50 heads. We would be quite astonished if all of the coins turned up heads. We would, a priori, (i.e., before any actual experience) state with confidence that such an outcome was *improbable*.

Notice that, being careful scientists, we did not say impossible. As a matter of fact, we realize that if the experiment were repeated over and over again an enormous number of times, eventually 100 heads would turn up simultaneously. We do accept the inevitability of an improbable event given enough trials. While 100 heads turning up at the same time might be regarded as a near miracle, biologists in general accept the more nearly miraculous events of evolution as purely fortuitous.

In these introductory comments we have introduced some new terms, or more likely, we have introduced some old words in a new light. We have not attempted any rigid definitions so far. For pedagogic reasons, whenever it is possible, precise definitions will be given as summaries after the sense of a discussion has made them useful.

Before these introductory comments are concluded, the meaning of the words *statistics* and *probability* will be considered. Statistics may be defined as the science and technique of gathering, analyzing, summarizing data, and estimating the probability of inferences from these data. We shall occasionally use the singular form, statistic, to refer to a value either observed or generated in a random manner.

Probability is a word that might better be used here without definition, as its meaning is still a matter of active philosophic debate. The word shall be used as if it meant the long run relative frequency of multiple or repeated events.

With statistics and probability defined adequately for present purposes, it might be useful at this point to indicate some of the things that statistical analysis is not and to point out the areas where our technical concept of probability does not apply. Experimenters should be reminded occasionally that statistical analysis is not a magic way of converting poor data into good data. Statistical analysis helps one deal with random variability produced by unknown, small causes, but it helps little, if at all, in dealing with variability due to poor or inadequate experimental technique.

In ordinary conversation we frequently use the words "probable" or "probably." For example, we might say that Peru and Guatemala will probably sign a treaty tomorrow. Notice that there is no way of giving a meaningful fractional or percentage value to this probability, since the situation is not representative of a multiple or repeated event. Another example of the way in which probability may be misinterpreted is as follows: suppose 100 students apply for admission to a graduate department which has only 25 openings. At first, we might think that a friend has a one in four chance of being accepted and might even make a bet giving odds one way or the other in the same one to four proportion. This would be ignoring the obvious fact that although the friend would either be admitted or not be admitted, the outcome would not be determined merely by random chance. Random chance would operate only on the probability of our winning bets if, in total ignorance of the factors determining entry, we made the same bet with many members of the group applying for admission. Insurance companies and bookmakers operate on this same principle of safety in numbers, but in a nonrepetitive instance or in the individual case, probability statements may have little or no valid meaning.

SUMMARY

Statistics may be defined as the science and technique of gathering, analyzing, and making inferences from data. These inferences are stated as probabilities. In this connection the term "probability" is used as if it meant the long run relative frequency of multiple or repeated events. The data subjected to statistical analysis consists of two types of variables: (1) continuous and (2) discontinuous. The former consist of measurements against a scale while the latter consist of data obtained by counting or enumerating discrete, indivisible units.

2
THE RELATIONSHIP OF PROBABILITY TO RANDOMIZATION

We usually have a clear definition in mind, which is readily understood by others, when using the words or expressions "pure chance," "probability," and "random." However, some difficulty is encountered when these ideas are examined as closely as is necessary for their utilization in making calculations. The most common example of pure chance is the toss of a coin. Whenever we prefer to let fate make a decision, we toss a coin. The first official act at the start of a football or baseball game is the tossing of a coin.

Generally, we accept the notion that heads will occur about as often as tails and thus that each side has an equal chance of being favored. At least we feel that "in the long run" the equality will hold. The important point to notice is the implied condition of a long run. A sequence or run of three or four heads in succession would not be surprising. During a trial of ten coin tosses performed while this paragraph was written, the following sequence occurred.

HTTTTHTTHT

However, our confidence is still unshaken, and we continue to expect the proportion of heads to tails to even out if the run is sufficiently prolonged. The basis for this confidence lies in the symmetry of the coin.

In contrast, consider expectations with a coin flipping machine constructed as shown in Figure 2-1. If the coin is always placed snugly against the back stop of the pedestal, if the flipper always delivers the same amount of force, and if the coin is always placed in the heads up position, we would expect the coin to show the same face each time it is tossed. Thus, if the coil came up tails on the first

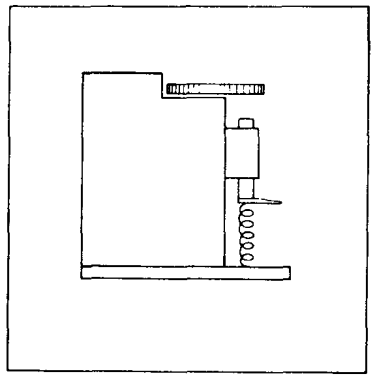


Figure 2-1. Mechanical coin tosser.

that, given the dimensions of the apparatus, the weight and diameter of the coin, and the force of the flipper, an engineer skilled in analytic mechanics could predict the outcome without any experimental trial at all.

Suppose, by way of contrast, we attempt to duplicate the machine's consistency by hand. We would immediately become aware of the many small but almost uncontrollable adjustments that would be necessary to make our flipping techniques competitive with the machine. We have already agreed that when the controlling factors are known, the outcome of an individual toss is known. Therefore, we must agree that the difference between a completely predictable outcome and a random outcome is our inability to control or evaluate the numerous, small factors that determine that outcome. In essence then, at least as far as coin tossing is concerned, the operative and determining factors in one toss may or may not be dominant in the next toss. With the machine the results show complete bias; when we try to imitate the machine by hand, incomplete but fairly substantial bias is introduced; when we simply toss the coin without any attempt to control it (i.e., at random), any remaining bias must be negligibly small.

Thus, *random* may be defined in a negative way as absence of bias or of factors known to contribute significantly to the outcome of any trial. To stress the more positive approach, we may consider a result to be random when the outcome is determined by the inconsistent interplay of many small factors.

A slightly more complicated but highly useful example is found in the classic white ball-black ball problems. Suppose we have in an urn 1000 balls, which are physically identical except that 100 of them are black and 900 are white. The container is shaken so that the balls move about at random. Theoretically, it would be possible to calculate the locations of each of these balls if all the starting positions, weights, force vectors, etc., were known. Obviously, this calculation would be too enormously complex to make it worthwhile; but the final location of each ball would be the result of the interplay of many small, virtually incalculable forces. In short, the balls would be distributed *randomly*.

Suppose further that ten balls are withdrawn by a blindfolded experimenter and the number of black balls in the sample recorded. The balls are returned to the container, shaken up (i.e., the balls are randomly redistributed) and the experiment repeated. This process is carried out over and over again. Just as in the case of the coin tossing experiment, we are able to predict *a priori* that in the long run, one out of every ten balls sampled will, on the average, be black. Similarly, remembering the definition at the end of Chapter I, we would say that the probability of selecting a black ball by chance from the container is one in ten, or more concisely, 1/10, as we would say of the tossed coin that the probability of heads turning up is 1/2.

To emphasize the lesson a little more strongly, consider, for example, a bag containing only five balls; one is white, and the other four are black. The experimenter shakes the bag without looking and draws a ball at random. Its color is noted, and the ball is returned to the bag. Repeating this procedure over and over again, we would say, *a priori,* that the probability of selecting a white ball by chance is 1/5. Thus, the principal element in predictions of this kind is not the large or small number of balls used; it is the large number of trials, or in other words, it is the *long run.*

The essential point of the foregoing examples is that we can state *a priori* the value of the pertinent probability when we know all the possibilities that can occur at random. In real life, particularly in the biological sciences, we rarely have such complete knowledge. Instead of predicting probabilities ahead of time from a knowledge of all the facts, we have the more difficult task of inferring some fragments of knowledge from observed proportions in the available data.